Transcription-associated compositional skews in Drosophila genes

Juraj Bergman[1,2], Andrea J. Betancourt[1,3], Claus Vogl*[,4]

[1]Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria

[2]Vienna Graduate School of Population Genetics, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria

[3]Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool, L69 7ZB, UK (current address)

[4]Institut für Tierzucht und Genetik, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria

*Corresponding author: E-mail: claus.vogl@vetmeduni.ac.at

Abstract

In many organisms, local deviations from Chargaff's second parity rule are observed around replication and transcription start sites and within intron sequences. Here, we use expression data as well as a whole-genome dataset of nearly 200 haplotypes to investigate such compositional skews in *Drosophila melanogaster* genes. We find a positive correlation between compositional skew and gene expression, comparable in strength to similar correlations between expression levels and genome-wide sequence features. This correlation is relatively stronger for germline, compared to somatic expression, consistent with the process of transcription-associated mutation bias. We also inferred

mutation rates from alleles segregating at low frequencies in short introns, and show that, while the overall GC content of short introns does not conform to the equilibrium expectation, the level of the observed deviation from the second parity rule is generally consistent with the inferred rates.

Key words: Chargaff's second parity rule, compositional skew, transcription-associated mutation bias, base composition evolution.

Introduction

Chargaff's second parity rule, *i.e.* the equal proportion of complementary nucleotide bases ([A]=[T] and [G]=[C]) along a strand of DNA, holds globally for most double-stranded DNA genomes (Mitchell and Bridge, 2006). Nevertheless, local deviations from this rule are common, especially around replication origins and transcription start sites and within introns (Francino and Ochman, 1997; Frank and Lobry, 1999; Touchon *et al.*, 2004). Compositional "skew" between strands may be introduced by DNA replication and transcription as a consequence of the directionality of DNA and RNA polymerization. Such skews have been regularly used to identify replication origins (oris) and termini in bacteria (*e.g.*, Lobry, 1996; Mrazek and Karlin, 1998; Picardeau *et al.*, 2000; Zawilak *et al.*, 2001). Recent technological advances in nascent strand purification allowed identification of oris in metazoans as well, and revealed similar skews surrounding these regions (Cayrou *et al.*, 2011, 2012; Comoglio *et al.*, 2015). Compositional skew in transcribed regions has also been observed, which has generally been attributed to transcription-associated mutation bias (TAMB) (Green *et al.*, 2003; McVicker and Green, 2010; Mugal *et al.*, 2009; Touchon *et al.*, 2003, 2004). TAMB might arise due to conditions differing between strands during transcription as one strand is chemically associated with the transcriptional machinery and the other exposed in the nucleus, which might result in strand-specific mutation or repair processes (Fong *et al.*, 2013; Svejstrup, 2002).

Here, we investigate compositional skews associated with transcription in *Drosophila melanogaster* using developmental and tissue-specific expression datasets (Chintapalli *et al.*, 2007; Graveley *et al.*, 2011; Vibranovski *et al.*, 2009) and sequence data from a large population sample of the ancestral range of the species (Lachaise *et al.*, 1988; Lack *et al.*, 2015). We find that non-coding regions within genes show strand-specific deviations from the second parity rule consistent with TAMB, while the overall GC content deviates from mutational equilibrium, as has been shown before (Kern and Begun, 2005; Zeng and Charlesworth, 2010; Clemente and Vogl, 2012; Robinson *et al.*, 2014).

Results

Association between compositional skews and gene expression

To investigate compositional skews between strands in transcribed regions, we calculated per gene estimates of CG skew (S_{CG}) and TA skew (S_{TA}) from coding-strand intron sequences, for the 1,925 autosomal and 478 X-linked genes that passed our data filtering (see Materials and Methods). If transcription helps to shape skews, this should be reflected in correlations between skews and gene expression. We thus examined correlations between skews and gene expression across different *D. melanogaster* tissues and developmental stages (Chintapalli *et al.*, 2007; Vibranovski *et al.*, 2009; Graveley *et al.*, 2011) for patterns consistent (or inconsistent) with TAMB.

The skew values calculated by concatenating all introns are S_{CG} = 1.18% (95% CI: 0.97%-1.37%) and S_{TA} = 0.82% (95% CI: 0.66%-0.97%). When looking at per gene skew estimates, we find that both S_{CG} and S_{TA} are positively, though weakly, correlated with gene expression (Fig. 1), consistent with TAMB, and in keeping with the known preference of C and T content on the coding strand of *Drosophila* introns (Touchon *et al.*, 2004). As expected, the skew parameters are also positively correlated with each other (Spearman's ρ=0.064, P=0.002), as has been observed for humans (Touchon *et al.*, 2003).

Secondly, spatial and temporal patterns of gene expression and skew are also broadly consistent with some effect of transcription. Specifically, only mutations occurring in the germline, or early in development (prior to the differentiation of germline tissues), are inherited and thus affect long-term base composition (McVicker and Green, 2010; Touchon *et al.*, 2003). Thus, we asked how the strengths of the correlations between skew and expression depend on expression in the germline or developmental stage. In fact, the correlation is relatively stronger between skew and expression levels in germ cells, for both ovaries and testes, than for somatic expression (Fig. 1A, D), as is also observed in humans (McVicker and Green, 2010) and in mice spermatogonia (Arneado *et al.*, 2011). Further, in a dataset consisting of gene expression for three different tissues of the *Drosophila* testes, the association between skew and expression during early spermatogenesis (in mitotic and meiotic cells) is stronger than that

between skew and post-meiotic expression (Fig. 1B, E). Similarly, the association between skew and gene expression is stronger for early developmental expression than for later developmental stages (Fig. 1C, F). All correlation coefficients are listed in Table S1.

While these trends are consistent with TAMB, there are a few caveats that require further analysis. The enrichment of C content on the coding strand could, in principle, be due to annotation errors if many annotated introns are in fact protein coding exons, as these tend to be C-rich in *Drosophila* (Akashi, 1994). However, the correlation patterns remain qualitatively similar when excluding introns with lengths that are multiples of three (Fig. S1), which are the most likely to be misannotated exons, as they do not imply frame-shifts. In addition, none of the comparisons based on germline, somatic or developmental stage expression individually show statistically significant differences (as indicated by the overlapping confidence intervals in Fig. 1), though all are stronger in the direction predicted by the TAMB hypothesis. Notably, the correlation between somatic expression and skew is positive (Fig. 1A, D), which is not a prediction of the TAMB hypothesis. However, this correlation may be a by-product of the positive correlation between somatic and germline expression. We therefore calculated pairwise partial correlations for skew values and gene expression for ovaries, testes and soma to estimate the independent effects of each. The results show that there is no relationship between skew values and somatic expression, while germline expression remains significantly correlated to skew (Tables S2 and S3). Furthermore, we compare the level of skews between concatenated introns from the 10% most highly and lowly expressed genes in ovaries and testes (Table 1). As expected, the 10% most highly expressed genes have significantly higher skew values compared to the 10% most lowly expressed genes, for both ovaries and testes (as indicated by the non-overlapping confidence intervals in Table 1). These results indicate that the level of skew is mainly driven by germline expression.

Population genetic analysis

To further analyze the likely causes of strand-specific nucleotide composition, we analyzed sites in autosomal short introns (≤65 bp in length) in a sample of

197 Zambian chromosomes from the putatively ancestral population of *D. melanogaster* from Zambia (Lack *et al.*, 2015; Table S4). Short introns appear to be the least selectively constrained of all sequence classes (Clemente and Vogl, 2012; Halligan and Keightley, 2006; Parsch *et al.*, 2010), and thus should most closely reflect mutational processes. We first focus on the sites that are fixed (*i.e.*, monomorphic) in the population sample alignment for one of the four possible nucleotides. The proportion of sites fixed for complementary nucleotides on the coding strand of autosomal short introns differs from a 1:1 ratio (calculated from a total of n = 191,747 sites; Table S4), with an excess of C over G (17.35% C *vs.* 14.91% G; χ^2 = 352.8, *d.f.* = 1, P<0.001) and T over A nucleotides (34.49% T *vs.* 33.24% A; χ^2 = 43.871, *d.f.* =1, P<0.001). Therefore, the resulting skews in short introns are S_{CG} = 7.55% (95% CI: 6.78%-8.34%) and S_{TA} = 1.84% (95% CI: 1.35%-2.41%). A similar pattern holds for all autosomal introns (n = 9,894,445 sites; 30.04% A, 19.87% C, 19.25% G, 30.84% T; Table S4), though the skew is weaker than for short introns — S_{CG} = 1.55% (95% CI: 1.46%-1.65%) and S_{TA} = 1.32% (95% CI: 1.24%-1.40%) — probably because the sequence composition of long introns is more selectively constrained (Haddrill *et al*, 2005). Notably, the G:C ratio in short introns (≈0.86) is more extreme (χ^2= 202.03, *d.f.* = 1, P<0.001) than the A:T ratio (≈0.96).

As an alternative to mutation, strand-specific selection could explain strand-specific skews. In particular, selection to avoid the canonical GT and AG splicing signals within intron sequences (Farlow *et al.*, 2012) might lead to an excess of C content on the coding strand compared to the non-coding strand (though note that short introns are AT-rich overall). To test whether mutation alone is sufficient to explain the observed compositional patterns, we estimated mutation rates from singleton frequencies of the autosomal short introns, *i.e.,* from sites in the sample alignment which contain a single copy of the minor allele variant. The relatively young age of these low frequency mutations makes it unlikely that their composition has been affected by directional selection (Kimura and Ohta, 1973; Messer, 2009); instead it should be predominantly influenced by mutation. The mutation rates estimated from singleton frequencies (Table 2) agree with previous estimates (Fig. S2). These rates indicate that any mutational asymmetry between coding and non-coding strands

is weak, similar to previous findings (*e.g.* Zeng, 2010). Furthermore, we find no evidence for mutation-associated compositional skews when directly comparing singleton frequencies between the coding and non-coding strands for each of the complementary nucleotide pairs (Table S5). When we analyse the estimated strand-specific mutation rates overall, however, they do imply C- and T-biased composition on the coding strand: the G-to-C and A-to-T rates are 1.1 and 1.08 times higher than their corresponding reverse mutation rate estimates. Given the point estimates of the mutation rates from Table 2, the equilibrium frequencies of fixed sites on the coding strand are (π_A, π_C, π_G, π_T)=(0.3653, 0.1245, 0.1231, 0.3871), resulting in the equilibriums skews of S_{CG} = 0.57% and S_{TA} = 2.90%. The resulting expected A:T ratio based on mutation rates deviates from a 1:1 ratio (χ^2 = 121.11, *d.f.* = 1, P<0.001) in the direction of the observed T over A excess, but more extremely (with an expected ratio equal to 1.060 *vs*. an observed ratio of 1.037; χ^2 = 14.592, *d.f.* = 1, P<0.001). The expected G:C ratio is in the same direction as the observed C-bias for the coding strand, but does not differ significantly from a 1:1 ratio (χ^2 = 1.518, *d.f.* = 1, P=0.218). Importantly, these results have been obtained using only point estimates, when in reality there is uncertainty in the estimates. We therefore used a parameter search algorithm (see Materials and Methods) and asked whether there are combinations of mutation rates, within the 95% confidence intervals of these estimates, which are consistent with the data. Specifically, we asked if mutation rates can explain both the skew and overall base composition. The results show that the estimated mutation rates can explain the levels of CG and TA skew observed in short intron sequences (Fig. 2A, B). However, the combinations of mutation rates that give rise to the observed levels of skew cannot explain the overall base composition in short introns (Fig. 2C, D) — the observed GC-content is too high to be consistent with these rates, *i.e.*, the GC content is not in mutational equilbrium, as noted by previous studies (Kern and Begun, 2005; Zeng and Charlesworth, 2010; Clemente and Vogl, 2012; Robinson *et al.*, 2014).

Finally, we used a generalized linear model (GLM) to analyze the association between mutation rates and gene expression. Specifically, we analyse the effect of expression on the frequency of singleton mutations from nucleotide *i*-to-*j* for the coding strand, using a GLM with a binomial response

variable consisting of successes (singletons of type *i*) and failures (fixed sites of type *j*); the expression estimates from ovaries, testes or soma (Chintapalli *et al*., 2007) were taken as explanatory variables. We analysed either singletons only in short introns (Table S6), and, in order to perform a more powerful analysis, we repeated the analysis using singletons in all introns (Table S7). As singletons are unlikely to be affected by selection (Kimura and Ohta, 1973; Messer, 2009), restricting this analysis to putatively neutral short introns may unnecessarily limit power. The results show that the correlations are, regardless of which dataset is used, consistently negative with few exceptions, suggesting a possible role of transcription-coupled repair in reducing overall mutation rates (Fong *et al.*, 2013; Svejstrup, 2002). In cases where the results of the GLM analyses indicate expression as a significant predictor of mutation rates, the associated coefficient is usually negative, implying that transcription is not mutagenic overall. Nonetheless, correlation coefficients associated with C or T singletons tend to be less negative than those associated with G or A singletons (Fig. S3), implying that mutation rates change with expression in a manner consistent with the observed directions of compositional skews.

Discussion

In eukaryotes, transcription appears to drive asymmetries in the frequencies of complementary nucleotides between the coding and non-coding strands of transcribed regions (McVicker and Green, 2010; Touchon *et al.*, 2003, 2004). Generally, T is more abundant than its complement A on the coding strand, while either G or C content is observed at higher frequencies in vertebrates or invertebrates, respectively (Touchon *et al.*, 2004).

In *D. melanogaster*, we find that gene expression in different tissues and across development correlates with compositional skew in a manner consistent with TAMB (Fig. 1). However, these correlations are weak and explain only a small proportion of the variance in skew levels between genes. The reason that the TAMB signal is weak is likely partly due to the fact that base composition in *Drosophila* introns changes with sequence length, and is affected by both purifying and positive selection (Singh *et al.*, 2005, Andolfatto, 2005; Haddrill *et*

al., 2005; Halligan and Keightley, 2006; Haddrill and Charlesworth, 2008). Nevertheless, weak genome-wide correlations can shed light on molecular processes shaping nucleotide base composition over evolutionary time: *e.g.,* the relationship between intronic GC content and recombination is similarly weak, but probably reflects the action of GC-biased gene conversion, now a well-established phenomenon of eukaryote genome evolution (Pessia *et al.*, 2012).

Materials and Methods

Data used in the analyses

Expression data were taken from Chintapalli *et al.* (2007), Vibranovski *et al.* (2009) and Graveley *et al.* (2011). The raw expression estimates were transformed with $\log_2$(value+1). For RNAseq data, these values are FPKM values; for the microarray analyses, they are relative flourescence intensities. Per gene expression values for soma and later developmental stages were calculated as averages across the non-germline and late developmental stage expression values, respectively. Replication start sites (RSS) were determined as peaks of the nascent strand signal or as a site of maximum coverage within a given ori region as identified in Cayrou *et al.* (2011) and Comoglio *et al.* (2015), respectively. We further analyzed a sample of the Zambian *D. melanogaster* population (Lack *et al.*, 2015). In total, the dataset consists of 197 sequences for each autosome and 196 sequences for the X chromosome. Sequences were annotated using the reference genome annotation of *D. melanogaster* (r5.57 from http://www.flybase.org/). For statistical analyses, R (R Core Team, 2014) was used.

Calculation of the skew parameters

The skew parameters (S_{CG} and S_{TA}) were calculated for each gene using the *D. melanogaster* reference sequence (r5.57 from http://www.flybase.org/). All intron sequences of the longest transcript of a gene were concatenated and estimates of skews per gene were calculated as: $S_{CG}=(C-G)/(C+G)$ and $S_{TA}=(T-A)/(T+A)$. Additionally, seven bases were trimmed from the 5' end and 35 bases from the 3' end of each intron to exclude genomic regions where the nucleotide

composition is affected by the presence of splicing sites (Fig. S4). Furthermore, genes overlapping regions ±500 bp around RSS were excluded from the analysis. Only genes containing ≥100 bp of concatenated intron sequence were considered. This filtering procedure left us with 1,925 autosomal and 478 X-linked genes available for analysis. The 95% confidence intervals for each of the skew parameters were estimated from 1,000 bootstrap-resamples. Each resample consisted of the number of observations equal to the number of sites used to calculate the original skew parameter, and the probabilities of drawing a specific nucleotide equal to the observed relative frequencies of nucleotides.

Inference of site frequency spectra and mutation rates

Site frequency spectra were inferred from the Zambian *D. melanogaster* sample (Lack *et al.*, 2015) for all six possible combinations of base pairs, for both autosomal short introns (≤65 bp in length; Clemente and Vogl, 2012; Halligan and Keightley, 2006; Parsch *et al.*, 2010), and all introns (Table S4). Using custom Python scripts, we filtered out sites that overlapped coding sequences or contained an undefined nucleotide state in at least one of the sequences in the sample alignment. Furthermore, sites belonging to the longest transcript of a gene were considered and sites with more than two alleles were filtered out. Intron sequences were trimmed as described previously. Mutation rates were calculated from autosomal short intron sequences as $q_{ij}=F_{ij}/M_i$, where q_{ij} indicates the mutation rate from nucleotide i to j, F_{ij} is the frequency of singletons of type j with major allele i, and M_i is the sum of the frequency of sites fixed for nucleotide i and the frequency of singletons of type F_{ij}. The confidence intervals for the mutation rate estimates were determined by assuming binomial probabilities with the number of successes $x=F_{ij}$ and the number of corresponding observations $n=M_i$.

Analysis of skew level and mutation rate estimates

We applied a parameter optimization algorithm to search for combinations of mutation rates, within their respective 95% confidence intervals (Table 1), which would recapitulate the observed levels of skew. To this end, we utilized

the Sequential Least SQuares Programming (SLSQP) method as implemented in the Python library "scipy" (Jones *et al.*, 2001). The parameters for each optimization run were randomly initialized within the 95% confidence intervals of the inferred mutation rates.

Generalized linear model analysis

The generalized linear model (GLM) analysis was conducted using the "glm" function in R (R Core Team, 2014) with the response variable following a binomial distribution and the default logit link function. The response variable was given as a two-column matrix where the first column contained the number of singletons of a specific type ("successes") while the second contained the number of corresponding fixed sites ("failures"). These frequencies were estimated per gene. The explanatory variables were gene expression estimates for either ovaries, testes or soma provided in Chintapalli *et al.* (2007).

Supplementary Material

Supplementary Tables S1-S7 and Figures S1-S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank all members of the Institute of Population Genetics for support and discussion. We also thank two anonymous reviewers whose suggestions helped to improve the manuscript. The work was funded by the Austrian Science Fund (FWF; grant number W1225-B20).

References

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 136(3):927-935.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437(7062), 1149-1152.

Arneodo A, Vaillant C, Audit B, Argoul F, d'Aubenton-Carafa Y, Thermes C. 2011. Multi-scale coding of genomic information: From DNA sequence to genome structure and function. Phys Rep. 498, 45-188.

Cayrou C, Coulombe P, Puy A, Rialle S, Kaplan N, Segal E, Mechali M. 2012. New insights into replication origin characteristics in metazoans. Cell Cycle 11(4):658-667.

Cayrou C, Coulombe P, Vigneron A, Stanojcic S, Ganier O, Pei er I, Rivals E, Puy A, Laurent-Chabalier S, Desprat R, Mechali M. 2011. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res. 21:1438-1449.

Chintapalli V, Wang J, Dow J. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet. 39:715-720.

Clemente F, Vogl C. 2012. Unconstrained evolution in short introns?-An analysis of genome-wide polymorphism and divergence data from Drosophila. J Evol Biol. 25(10):1975-1990.

Comoglio F, Schlumpf T, Schmid V, Rohs R, Beisel C, Paro R. 2015. High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. Cell Rep. 11:821-834.

Farlow A, Dolezal M, Hua L, Schlötterer C. 2012. The genomic signature of splicing-coupled selection differs between long and short introns. Mol Biol Evol. 29(1):21-24.

Fong YW, Cattoglio C, Tjian R. 2013. The intertwined roles of transcription and repair proteins. Mol Cell 52(3):291-302.

Francino PM, Ochman H. 1997. Strand asymmetries in DNA evolution. Trends Genet. 13(6):240-245.

Frank AC, Lobry JR. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238(1):65-77.

Graveley BR, Brooks AN, Carlson JW, Du MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE. 2011, Mar. The developmental transcriptome of Drosophila melanogaster. Nature 471(7339):473-479.

Green P, Ewing B, Miller W, Thomas PJ, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet. 33(4):514-517.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. Genome Biol. 6(8):R67.

Haddrill PR, Charlesworth B. 2008. Non-neutral processes drive the nucleotide composition of non-coding sequences in Drosophila. Biol Lett. 4(4):438-441.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. Genome Res. 16(7):875-884.

Jones E, Oliphant T, Peterson P, et al. 2001. SciPy: Open source scientific tools for Python. [Online; accessed 2016-11-30].

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. Genome Res. 19(7):1195-1201.

Kern AD, Begun DJ. 2005. Patterns of polymorphism and divergence from noncoding sequences of Drosophila melanogaster and D. simulans: evidence for nonequilibrium processes. Mol Biol Evol. 22:51-62.

Kimura M, Ohta T. 1973. The age of a neutral mutant persisting in a finite population. Genetics 75:199-212.

Lachaise D, Cariou ML, David JR, Lemeunier F, Tsacas L, Ashburner M. 1988. Historical biogeography of the Drosophila melanogaster species subgroup. Evol Biol. 22:159-225.

Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE. 2015. The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. Genetics 199(4):1229-1241.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 10:e166.

Lobry J. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol. 13:660-665.

Marais G, Mouchiroud D, Duret L. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A 98(10):5688-5692.

McVicker G, Green P. 2010. Genomic signatures of germline gene expression. Genome Res. 20:1503-1511.

Messer PW. 2009. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. Genetics 182(4):1219-1232.

Mitchell D, Bridge R. 2006. A test of Chargaff's second rule. Biochem Bioph Res Co. 340:90-94.

Mrazek J, Karlin S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci U S A 95(7):3720-3725.

Mugal CF, von Grünberg HH, Peifer M. 2009. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. Mol Biol Evol. 26(1):131-142.

Parsch J, Novozhilov S, Saminadin-Peter S, Wong K, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. Mol Biol Evol. 27(6):1226-1234.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais G. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biol Evol. 4(7):675-682.

Picardeau M, Lobry JR, Hinnebusch BJ. 2000. Analyzing DNA strand compositional asymmetry to identify candidate replication origins of Borrelia burgdorferi linear and circular plasmids. Genome Res. 10(10):1594-1604.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ, Langley CH. 2012. Population genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genet. 8(12):e1003080.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Robinson MC, Stone EA, Singh ND. 2014. Population genomic analysis reveals no evidence for GC-biased gene conversion in Drosophila melanogaster. Mol Biol Evol. 31(2):425-433.

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. Genetics 194(4):937-954.

Sharp NP, Agrawal AF. 2016. Low genetic quality alters key dimensions of the mutational spectrum. PLoS Biol. 14(3):e1002419.

Singh ND, Davis JC, Petrov DA. (2005). Codon bias and noncoding GC content correlate negatively with recombination rate on the Drosophila X chromosome. J Mol Evol. 61:315-324.

Svejstrup JQ. 2002. Mechanisms of transcription-coupled DNA repair. Nat Rev Mol Cell Biol. 3(1):21-29.

Touchon M, Arneodo A, d'Aubenton Carafa Y, Thermes C. 2004. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. Nucleic Acids Res. 32(17):4969-4978.

Touchon M, Nicolay S, Arneodo A, d'Aubenton-Carafa Y, Thermes C. 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. FEBS Lett. 55:579-582.

Vibranovski M, Lopes H, Karr T, Long M. 2009. Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. PLoS Genet. 5(11).

Zawilak A, Cebrat S, Mackiewicz P, Krol-Hulewicz A, Jakimowicz D, Messer W, Gosciniak G, Zakrzewska-Czerwinska J. 2001. Identification of a putative

chromosomal replication origin from Helicobacter pylori and its interaction with the initiator protein DnaA. Nucleic Acids Res. 29(11):2251-2259.

Zeng K, Charlesworth B. 2010. Studying patterns of recent evolution at synonymous sites and intronic sites in Drosophila melanogaster. J Mol Evol. 70:116-128.

Zeng K. 2010. A simple multiallele model and its application to preferred-unpreferred codons using polymorphism data. Mol Biol Evol. 27:1327-1337.

TABLE 1. – Skew values with their corresponding 95% confidence intervals (in square brackets) calculated from concatenating introns in genes with the 10% highest and 10% lowest expression in different germline tissues (number of genes in each category is n = 141).

	Ovary expression		Testes expression	
	high	low	high	low
S_{CG} (%)	4.19 [3.24, 5.17]	0.67 [-0.04, 1.36]	4.33 [3.18, 5.39]	0.55 [-0.08, 1.19]
S_{TA} (%)	2.31 [1.54, 3.06]	-0.45 [-0.99, 0.15]	2.58 [1.75, 3.39]	-0.16 [-0.68, 0.35]

TABLE 2. - Mutation rates q_{ij} from nucleotide i to j with the corresponding 95% confidence intervals, estimated from the coding strand of autosomal short introns.

$i \rightarrow j$	q_{ij} (F_{ij}/M_i)	q_{ij} 95% CI
A → C	0.0048 (308/64,053)	0.0043-0.0053
A → G	0.0120 (776/64,521)	0.0112-0.0128
A → T	0.0110 (709/64,454)	0.0102-0.0118
C → A	0.0181 (615/33,886)	0.0167-0.0195
C → G	0.0080 (270/33,541)	0.0071-0.0089
C → T	0.0293 (1,008/34,359)	0.0276-0.0310
G → A	0.0324 (957/29,556)	0.0304-0.0344
G → C	0.0089 (256/28,885)	0.0079-0.0099
G → T	0.0175 (508/29,107)	0.0161-0.0190
T → A	0.0100 (667/66,779)	0.0093-0.0107
T → C	0.0113 (753/66,885)	0.0105-0.0121
T → G	0.0047 (314/66,446)	0.0042-0.0052

NOTE. – F_{ij} is the frequency of singletons of type j with major allele i and M_i is the sum of the frequency of sites fixed for nucleotide i and the frequency of singletons of type F_{ij}.

Figure Legends

FIG. 1. - Pearson's coefficients (with 95 % confidence intervals) for the correlations between compositional skew and gene expression across different tissues and developmental stages. (A-C) Correlation of CG skew and gene expression. (D-F) Correlation of TA skew and gene expression. While 0-2 hour expression in embryos most likely reflects maternal transcription, which should not necessarily affect germline development, the correlation between maternal expression and later putative zygotic expression (2-4 hour) is strong (Spearman's ρ=0.937, P<0.001).

FIG. 2. - Distributions of skew estimates and nucleotide content obtained from 10,000 independent parameter search runs, conditional on the observed compositional skew in autosomal short introns and the 95% confidence intervals of mutation rates in Table 1. (A-B) Distributions of CG and TA skew, respectively; the red dashed line is the observed skew level. (C) The distribution of G (red) and C (black) content; the dashed lines are the observed values. (D) The distribution of A (red) and T (black) content; the dashed lines are the observed values.

Figures

FIG. 1.

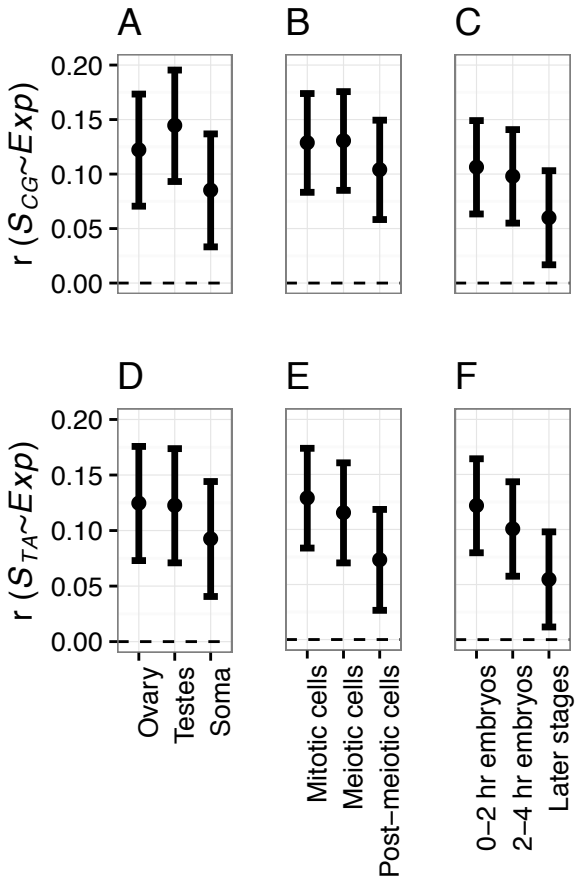

FIG. 2.